

Semi-supervised logistic discrimination via labeled data and unlabeled data from different sampling distributions

Shuichi Kawano

*Department of Mathematical Sciences, Graduate School of Engineering,
Osaka Prefecture University, 1-1 Gakuen-cho, Sakai, Osaka 599-8531, Japan.*

skawano@ms.osakafu-u.ac.jp

Abstract: This article addresses the problem of classification method based on both labeled and unlabeled data, where we assume that a density function for labeled data is different from that for unlabeled data. We propose a semi-supervised logistic regression model for classification problem along with the technique of covariate shift adaptation. Unknown parameters involved in proposed models are estimated by regularization with EM algorithm. A crucial issue in modeling process is the choices of tuning parameters in our semi-supervised logistic models. In order to select the parameters, a model selection criterion is derived from information-theoretic approach. Some numerical studies show that our modeling procedure performs well in various cases.

Key Words and Phrases: Covariate shift, EM algorithm, Model selection, Regularization, Semi-supervised learning.

1 Introduction

In recent years, with the wide availability of fast and high-powered computers, high-throughput data of unexampled size and complexity have frequently been seen in contemporary statistics and machine learning. Examples involve data from genomics, proteomics, natural language processing, and signal processing. For the huge amount of data, it is difficult to label data by human operator, since its work requires vast times and efforts. Only small labeled data set may, therefore, be available, while unlabeled data set can

be more easily obtained. Under such a circumstance, a classification method that combines both labeled and unlabeled data, called semi-supervised learning, has received an enormous amount of attention in the late machine learning and statistical literature (see, e.g., Chapelle *et al.*, 2006; Liang *et al.*, 2007). For overviews of semi-supervised learning methods, we refer to Zhu (2008), and references given therein.

Many classification techniques for semi-supervised learning have been proposed by various researchers, e.g., Amini and Gallinari (2002), Basu *et al.* (2004), Bennett and Demiriz (1998), Chen and Wang (2007), Dean *et al.* (2006), Kawano *et al.* (2010), Kawano and Konishi (2011), Lafferty and Wasserman (2007), and Zhou *et al.* (2004). Most of these semi-supervised methods implicitly assumes that a density function for labeled data is the same as that for unlabeled data. On the other hand, we, here, consider the case that the densities for labeled data and unlabeled data are different, since the densities are not always same in practical situations. In such a case, several semi-supervised methods have been presented, e.g., Jiang and Zhai (2007), Wu *et al.* (2009), and Zadrozny (2004). However, for these methods, there remains a problem of evaluating constructed semi-supervised models, which is a crucial issue in model building process. Cross validation (CV) is often used in evaluating models constructed by semi-supervised procedure. An advantage of CV lies in its independence from probabilistic assumptions. The computational time of the procedures is, however, very large, and the high variability and tendency to undersmooth in CV are not negligible in the analysis of complex or high-dimensional data, since the selectors are repeatedly applied.

In this paper, we propose a logistic model for semi-supervised classification problem by using statistical methods under covariate shift (Shimodaira, 2000) in the case that the density function for labeled data is different from that for unlabeled data. The unknown parameters in the model are estimated by regularization method with the help of EM algorithm. A crucial issue in our modeling strategy is to choose values of some tuning parameters included in semi-supervised logistic models, which corresponds to evaluating models determined by our proposed procedures. In order to objectively select optimal values of tuning parameters, we then introduce a model selection criterion based on information-

theoretic approach (Konishi and Kitagawa, 1996) that evaluates semi-supervised logistic models estimated by regularization method. Some numerical examples demonstrate that the proposed procedure works well and performs better than competing methods.

This paper is organized as follows. In Section 2, we present a semi-supervised logistic model for classification problem based on covariate shift adaptation and its estimation by regularization method. Section 3 provides a model selection criterion derived from information-theoretic viewpoint to select some tuning parameters in our logistic models. In Section 5, Monte Carlo simulations and benchmark data analysis are given to assess the performances of proposed semi-supervised logistic discrimination. Some concluding remarks are given in Section 5.

2 Semi-supervised logistic modeling from different sampling distributions

2.1 Linear logistic modeling for semi-supervised learning

We review here semi-supervised linear logistic models developed by early researchers (e.g., Amini and Gallinari, 2002; Vittaut *et al.*, 2002). Suppose that we have an n_1 labeled data set $\{(\mathbf{x}_\alpha, y_\alpha); \alpha = 1, \dots, n_1\}$ and an $(n - n_1)$ unlabeled data set $\{\mathbf{x}_\alpha; \alpha = n_1 + 1, \dots, n\}$, where $\mathbf{x}_\alpha = (x_{\alpha 1}, \dots, x_{\alpha p})^T$ denotes a p -dimensional explanatory variable and Y_α is a random variable taking values 0 or 1 with probabilities

$$\Pr(Y_\alpha = 1|\mathbf{x}_\alpha) = \pi(\mathbf{x}_\alpha), \quad \Pr(Y_\alpha = 0|\mathbf{x}_\alpha) = 1 - \pi(\mathbf{x}_\alpha). \quad (1)$$

Note that logistic models are first constructed by only the labeled data set, while the unlabeled data set is used in estimating the parameters involved in the logistic models.

Using posterior probabilities in Equation (1) and the labeled data set, a linear logistic model (see, e.g., Hastie *et al.*, 2009) is formulated by

$$\log \left\{ \frac{\pi(\mathbf{x}_\alpha)}{1 - \pi(\mathbf{x}_\alpha)} \right\} = w_0 + \sum_{j=1}^p w_j x_{\alpha j} = \mathbf{w}^T \mathbf{x}_\alpha^*, \quad \alpha = 1, \dots, n_1, \quad (2)$$

where $\mathbf{w} = (w_0, w_1, \dots, w_p)^T$ is an unknown parameter vector and $\mathbf{x}_\alpha^* = (1, \mathbf{x}_\alpha^T)^T$. Hereafter, we denote posterior probabilities by $\pi(\mathbf{x}_\alpha; \mathbf{w})$, since the posterior probabilities depend on the parameter vector \mathbf{w} . It follows from Equation (2) that posterior probabilities can be rewritten as

$$\pi(\mathbf{x}_\alpha; \mathbf{w}) = \frac{\exp(\mathbf{w}^T \mathbf{x}_\alpha^*)}{1 + \exp(\mathbf{w}^T \mathbf{x}_\alpha^*)}. \quad (3)$$

Also, a probability function of random variable Y_α is the Bernoulli distribution in the form

$$f(y_\alpha | \mathbf{x}_\alpha; \mathbf{w}) = \pi(\mathbf{x}_\alpha; \mathbf{w})^{y_\alpha} \{1 - \pi(\mathbf{x}_\alpha; \mathbf{w})\}^{1-y_\alpha}, \quad y_\alpha = 0, 1. \quad (4)$$

Under the linear logistic model, the log-likelihood for y_α in terms of \mathbf{w} is induced into

$$\begin{aligned} \ell(\mathbf{w}) &= \sum_{\alpha=1}^{n_1} \log f(y_\alpha | \mathbf{x}_\alpha; \mathbf{w}) \\ &= \sum_{\alpha=1}^{n_1} [y_\alpha \log \pi(\mathbf{x}_\alpha; \mathbf{w}) + (1 - y_\alpha) \log \{1 - \pi(\mathbf{x}_\alpha; \mathbf{w})\}] \\ &= \sum_{\alpha=1}^{n_1} [y_\alpha \mathbf{w}^T \mathbf{x}_\alpha^* - \log \{1 + \exp(\mathbf{w}^T \mathbf{x}_\alpha^*)\}]. \end{aligned} \quad (5)$$

By ordinary, the unknown parameter \mathbf{w} included in the logistic model is estimated by maximizing the log-likelihood function with respect to the parameter. The procedure is known as the supervised learning, i.e., the parameter is determined by using only labeled data set. Since we have an additional unlabeled data set, the parameter should be estimated by both labeled and unlabeled data set, which is called the semi-supervised learning. Thereby, Amini and Gallinari (2002) proposed a log-likelihood function with additional unlabeled data given by

$$\begin{aligned} \ell^*(\mathbf{w}) &= \sum_{\alpha=1}^{n_1} [y_\alpha \mathbf{w}^T \mathbf{x}_\alpha^* - \log \{1 + \exp(\mathbf{w}^T \mathbf{x}_\alpha^*)\}] \\ &\quad + \sum_{\alpha=n_1+1}^n [t_\alpha \mathbf{w}^T \mathbf{x}_\alpha^* - \log \{1 + \exp(\mathbf{w}^T \mathbf{x}_\alpha^*)\}], \end{aligned} \quad (6)$$

where t_α ($\alpha = n_1 + 1, \dots, n$) is a latent variable coded as 0 or 1. Amini and Gallinari (2002) estimated the parameter by maximizing the Equation (6) with the technique of

EM algorithm, while Kawano and Konishi (2011) employed the Equation (6) with regularization term in estimating the parameter in the context of nonlinear logistic model based on basis expansion.

Given the estimate $\hat{\mathbf{w}}$, we assign a future observation \mathbf{x}_f into class j ($j = 0, 1$) that has the maximum posterior probability in the Equation (3).

2.2 Semi-supervised logistic model from different distributions

Logistic models using semi-supervised learning described in Section 2.1 usually assumes that a density function for the labeled data set is the same as that for the unlabeled data set, i.e., when we denote that $q_{\text{label}}(\mathbf{x})$ is a probability distributional function of explanatory variables for the labeled data and $q_{\text{unlabel}}(\mathbf{x})$ is that for the unlabeled data, $q_{\text{label}}(\mathbf{x}) = q_{\text{unlabel}}(\mathbf{x})$. Our aim in this section is to construct logistic models under the situation that a density for the labeled data set is different from that for the unlabeled data set, i.e., $q_{\text{label}}(\mathbf{x}) \neq q_{\text{unlabel}}(\mathbf{x})$.

We recall the log-likelihood function for logistic model with unlabeled data in Equation (6). For the log-likelihood function, we propose a weighted log-likelihood function with unlabeled data in the form

$$\begin{aligned} \ell^*(\mathbf{w}; \gamma_1, \gamma_2) = & \sum_{\alpha=1}^{n_1} \left\{ \frac{q_{\text{unlabel}}(\mathbf{x}_\alpha)}{q_{\text{label}}(\mathbf{x}_\alpha)} \right\}^{\gamma_1} [y_\alpha \mathbf{w}^T \mathbf{x}_\alpha^* - \log\{1 + \exp(\mathbf{w}^T \mathbf{x}_\alpha^*)\}] \\ & + \sum_{\alpha=n_1+1}^n \left\{ \frac{q_{\text{label}}(\mathbf{x}_\alpha)}{q_{\text{unlabel}}(\mathbf{x}_\alpha)} \right\}^{\gamma_2} [t_\alpha \mathbf{w}^T \mathbf{x}_\alpha^* - \log\{1 + \exp(\mathbf{w}^T \mathbf{x}_\alpha^*)\}] , \quad (7) \end{aligned}$$

where $\gamma_1, \gamma_2 \in [0, 1]$ are tuning parameters. If both γ_1 and γ_2 is 0, the log-likelihood in Equation (7) coincides with that in Equation (6). Note that the weight on the first term, $q_{\text{unlabel}}(\mathbf{x})/q_{\text{label}}(\mathbf{x})$, is bigger near high density of unlabeled data, while that on the second term, $q_{\text{label}}(\mathbf{x})/q_{\text{unlabel}}(\mathbf{x})$, is strengthen near high density of labeled data. Hence, the log-likelihood function on the first term is highly weighted near high density of unlabeled data, while that on the second term has high weighting near high density of labeled data. An idea of the weight, the ratio of $q_{\text{label}}(\mathbf{x})$ and $q_{\text{unlabel}}(\mathbf{x})$, arises from a statistical inference under covariate shift (Shimodaira, 2000). In semi-supervised learning, employing a ratio of densities in log-likelihood functions is not new. For example, Sokolovska *et al.* (2008)

and Zou *et al.* (2007) use a ratio of densities in semi-supervised inference. However, the Equation (7) is a novel formulation in semi-supervised context.

The Equation (7) includes unknown values of ratios, $q_{\text{unlabel}}(\mathbf{x})/q_{\text{label}}(\mathbf{x})$ and $q_{\text{label}}(\mathbf{x})/q_{\text{unlabel}}(\mathbf{x})$, which are to be estimated. Various researchers address the problem of estimating the ratios by using several methods of statistics or machine learning (Bickel *et al.*, 2009; Huang *et al.*, 2007; Kanamori *et al.*, 2009; Sugiyama *et al.*, 2008). In this paper, we employ a uLSIF method proposed by Kanamori *et al.* (2009) in determining values of the ratios and implement the method by a source code given in <http://www.math.cm.is.nagoya-u.ac.jp/~kanamori/software/LSIF>. We do not follow details of density ratio estimation by the uLSIF method, since these are not our focus in this paper. For readers that are interested in the topics, we refer to Kanamori *et al.* (2009).

2.3 Parameter estimation via regularization

In estimating parameters in logistic models, the log-likelihood function often diverges to infinity when the maximum likelihood method is applied (Konishi and Kitagawa, 2008). Hence, the parameter vector \mathbf{w} in Equation (7) is estimated by regularization method. The regularization method achieves to maximize a following regularized log-likelihood function

$$\ell_{\lambda}^*(\mathbf{w}; \gamma_1, \gamma_2) = \ell^*(\mathbf{w}; \gamma_1, \gamma_2) - \frac{n_1 \lambda}{2} \mathbf{w}^T K \mathbf{w}, \quad (8)$$

where λ is a regularization parameter that has positive values and $K = \text{diag}(0, I_p)$ is a $(p+1) \times (p+1)$ matrix. Here, the matrix I_p is a p -dimensional identity matrix.

It is not easy to optimize the parameter involved in Equation (8), since the latent variables t_{α} ($\alpha = n_1 + 1, \dots, n$) are unobserved. Hence, we employ an EM-based algorithm developed by Kawano and Konishi (2011) as follows:

Step1 Estimate the parameter vector \mathbf{w} by maximizing the regularized log-likelihood function using only labeled data set $\{(\mathbf{x}_{\alpha}, y_{\alpha}); \alpha = 1, \dots, n_1\}$ along with the technique of Newton-Raphson method.

Step2 Construct a classification rule $\pi(\mathbf{x}_{\alpha}; \hat{\mathbf{w}})$.

Step3 According to the classification rule in Step2, compute the posterior probabilities $\pi(\mathbf{x}_\alpha; \hat{\mathbf{w}})$ for unlabeled data set \mathbf{x}_α ($\alpha = n_1 + 1, \dots, n$). By the use of the posterior probabilities, estimate t_α in the form $\hat{t}_\alpha = \pi(\mathbf{x}_\alpha; \hat{\mathbf{w}})$.

Step4 Replace t_α into \hat{t}_α in the regularized log-likelihood function (8), and then determine the parameter vector \mathbf{w} through the maximization of the log-likelihood function in Equation (8) with the help of Newton-Raphson method.

Step5 Repeat the Step2 to the Step4 until the following condition

$$|\ell_\lambda^*(\hat{\mathbf{w}}^{(k+1)}; \gamma_1, \gamma_2) - \ell_\lambda^*(\hat{\mathbf{w}}^{(k)}; \gamma_1, \gamma_2)| < \varepsilon \quad (9)$$

is satisfied, where $\hat{\mathbf{w}}^{(k)}$ is the value of \mathbf{w} after the k -th EM iteration and ε is an arbitrary small number (e.g., 10^{-5}).

It follows from these procedures that we obtain a statistical model in the form

$$f(y|\mathbf{x}; \hat{\mathbf{w}}) = \pi(\mathbf{x}; \hat{\mathbf{w}})^y \{1 - \pi(\mathbf{x}; \hat{\mathbf{w}})\}^{1-y}. \quad (10)$$

Note that the statistical model is constructed by using both labeled data and unlabeled data.

3 Model selection criterion

The statistical model in Equation (10) contains some adjusted parameters including two tuning parameters γ_1, γ_2 in the weighted log-likelihood function and the regularization parameter λ . Regarding selection of these adjusted parameters as that of candidate models, we introduce a model selection criterion from information-theoretic approach.

Akaike (1974) introduced the Akaike information criterion (AIC) for evaluating statistical models estimated by maximum likelihood method. It is, however, difficult for the AIC to evaluate models given by estimation procedures except for maximum likelihood method, whereas the AIC is widely used in many fields of research. By extending the AIC, Konishi and Kitagawa (1996) derived an information criterion, which can evaluate models given by the M-estimator including regularization method. Using this result, we present

a generalized information criterion (GIC) for evaluating our proposed semi-supervised logistic models estimated by regularization method. The model selection criterion is given by

$$\text{GIC} = -2 \sum_{\alpha=1}^{n_1} \left\{ \frac{q_{\text{unlabel}}(\mathbf{x}_\alpha)}{q_{\text{label}}(\mathbf{x}_\alpha)} \right\}^{\gamma_1} \log f(y_\alpha | \mathbf{x}_\alpha; \hat{\mathbf{w}}) + 2\text{tr} \{ Q(\hat{\mathbf{w}}) R^{-1}(\hat{\mathbf{w}}) \}, \quad (11)$$

where the matrices $Q(\hat{\mathbf{w}})$ and $R(\hat{\mathbf{w}})$ are

$$Q(\hat{\mathbf{w}}) = \frac{1}{n_1} \left\{ X^T \hat{W}^2 \hat{\Lambda}^2 X - \lambda K \hat{\mathbf{w}} \mathbf{1}_{n_1}^T \hat{W} \hat{\Lambda} X \right\}, \quad (12)$$

$$R(\hat{\mathbf{w}}) = \frac{1}{n_1} X \hat{\Pi} \hat{W} (I_{n_1} - \hat{\Pi}) X + \lambda K. \quad (13)$$

Here, $\mathbf{1}_{n_1}$ is an n_1 -dimensional vector the elements of which are all 1, I_{n_1} is an n_1 -dimensional identity matrix. Also, X , \hat{W} , $\hat{\Lambda}$, and $\hat{\Pi}$ are, respectively, given by

$$\begin{aligned} X &= (\mathbf{x}_1^*, \dots, \mathbf{x}_{n_1}^*)^T, \\ \hat{W} &= \text{diag} \left[\left\{ \frac{q_{\text{unlabel}}(\mathbf{x}_1)}{q_{\text{label}}(\mathbf{x}_1)} \right\}^{\gamma_1}, \dots, \left\{ \frac{q_{\text{unlabel}}(\mathbf{x}_{n_1})}{q_{\text{label}}(\mathbf{x}_{n_1})} \right\}^{\gamma_1} \right], \\ \hat{\Lambda} &= \text{diag} [y_1 - \pi(\mathbf{x}_1; \hat{\mathbf{w}}), \dots, y_{n_1} - \pi(\mathbf{x}_{n_1}; \hat{\mathbf{w}})], \\ \hat{\Pi} &= \text{diag} [\pi(\mathbf{x}_1; \hat{\mathbf{w}}), \dots, \pi(\mathbf{x}_{n_1}; \hat{\mathbf{w}})]. \end{aligned}$$

We choose adjusted parameters from the minimizer of the GIC in Equation (11).

4 Numerical study

We studied some numerical examples to show the efficiency of our proposed modeling strategy. Two types of Monte Carlo simulations and benchmark data analysis are given to illustrate the proposed semi-supervised logistic discrimination.

4.1 Simulation 1

We investigated the effectiveness of the proposed modeling procedures through Monte Carlo simulation. In this simulation study, we generated data sets $\{(x_{1\alpha}, x_{2\alpha}, y_\alpha); \alpha = 1, \dots, n\}$ as labeled data and $\{(x_{1\alpha}, x_{2\alpha}); \alpha = 1, \dots, 500\}$ as unlabeled data. In labeled

Table 1: Comparisons of prediction error rates (%) for several number of data points.

Method \ # of labeled data	25	50	100	150	200	250
SSLRCS	33.3	33.3	33.9	34.8	35.5	35.0
LSSLR	34.3	34.4	34.2	35.3	35.9	35.6
SLR	35.6	34.3	34.3	35.2	35.8	35.6

data, $(x_{1\alpha}, x_{2\alpha})$ were generated by normal distribution $N((-0.9, 1 - \sin(\sin(0.9^2\pi)))^T, \text{diag}(0.0015, 2))$, and y_α was generated according to a following conditional probability

$$\Pr(Y = 1|x_1, x_2) = 1 / [1 + \exp \{-\sin(2\pi x_1^2) - x_2 + 1\}]. \quad (14)$$

Meanwhile, unlabeled data $(x_{1\alpha}, x_{2\alpha})$ were obtained by normal distribution $N((-0.4, 1 - \sin(\sin(0.4^2\pi)))^T, \text{diag}(0.05, 1))$. Test data set $\{(x_{1\alpha}, x_{2\alpha}, y_\alpha); \alpha = 1, \dots, 1000\}$ was generated as follows. First, $(x_{1\alpha}, x_{2\alpha})$ were derived by mixture of labeled and unlabeled data, where the mixing rate is equal (that is, 0.5). Second, for the $(x_{1\alpha}, x_{2\alpha})$, y_α was obtained according to the conditional probability in Equation (14). We assumed that labeled data sizes (n) were 25, 50, 100, 150, 200, and 250.

We fitted our semi-supervised logistic regression model to the data sets. Note that the density ratio estimation procedure by uLSIF method described in Section 2.2 is not performed in this simulation trials, since the density ratio is exactly calculated. The simulation results were obtained by averaging over 50 repeated Monte Carlo trial. The tuning parameters in our models were selected by using the GIC in Equation (11). The values of tuning parameters were $\gamma_1 = 0.10$, $\gamma_2 = 0.610$, and $\lambda = 10^{-2.20}$, which were averaged over 50 repetitions. The results are summarized in Table 1.

We compared the performances of the proposed semi-supervised methodologies (SSLRCS: semi-supervised logistic regression under covariate shift) with those of semi-supervised method proposed by Amini and Gallinari (2002) (LSSLR: linear semi-supervised logistic regression), which is developed under the condition that density functions for labeled and unlabeled data are same, and supervised linear logistic discriminant analysis (SLR: supervised logistic regression). Note that the SLR is constructed by using only labeled

data. Semi-supervised and supervised logistic modeling strategies were applied into the data sets. Since the LSSLR and the SLR include a tuning parameter, respectively, the parameter is determined by the GIC, where the GIC for LSSLR is obtained by setting $q_{\text{unlabel}}(\mathbf{x}_\alpha)/q_{\text{label}}(\mathbf{x}_\alpha) = 1$ ($\alpha = 1, \dots, n_1$) in Equation (11) and that for SLR is given by Ando *et al.* (2008). It may be seen from Table 1 that SSLRCS is superior to other methods (LSSLR and SLR) in all cases in the sense that the proposed method gives smaller prediction error rates.

4.2 Simulation 2

We simulated three data sets given in Chakraborty (2011) to examine the performances of our proposed modeling strategy. For each of the simulation cases, we generated 100 data points in the labeled data set, 1000 data points in the unlabeled data set, and 1000 data points in the test data set. Using the data sets, we constructed the SSLRCS, the LSSLR, and the SLR. We repeated the procedure 50 times. Our simulation settings are given as follows (for details, see, Chakraborty (2011, p. 76)):

- Case 1 : In the labeled data set, generate $\mathbf{x} = (x_1, x_2)^T$ given by $x_i \sim N(2, 1)$ ($i = 1, 2$) for Class 1 and $x_i \sim N(-2, 1)$ ($i = 1, 2$) for Class 2. In the unlabeled data set, $x_i \sim N(2, 2)$ ($i = 1, 2$) for Class 1 and $x_i \sim N(-2, 2)$ ($i = 1, 2$) for Class 2. In the test data set, $x_i \sim 0.5N(2, 1) + 0.5N(2, 2)$ ($i = 1, 2$) for Class 1 and $x_i \sim 0.5N(-2, 1) + 0.5N(-2, 2)$ ($i = 1, 2$) for Class 2.
- Case 2 : Generate $\mathbf{x} = (x_1, \dots, x_{10})^T$ given by $x_i \sim N(1, 3)$ ($i = 1, \dots, 10$) for Class 1 and $x_i \sim N(-1, 3)$ ($i = 1, \dots, 10$) for Class 2.
- Case 3 : Generate $\mathbf{x} = (x_1, x_2)^T$ given by $x_i \sim N(5, 2)$ ($i = 1, 2$) for Class 1 and $x_i \sim N(8, 2)$ ($i = 1, 2$) for Class 2 in the labeled data set. In the unlabeled data set, $x_i \sim N(6, 2)$ ($i = 1, 2$) for Class 1 and $x_i \sim N(9, 2)$ ($i = 1, 2$) for Class 2. In the test data set, $x_i \sim 0.5N(5, 2) + 0.5N(6, 2)$ ($i = 1, 2$) for Class 1 and $x_i \sim 0.5N(8, 2) + 0.5N(9, 2)$ ($i = 1, 2$) for Class 2.

Table 2: Comparisons of prediction error rates (%) for several cases.

Method \ Data sets	Case 1	Case 2	Case 3
SSLRCS	1.28	3.65	9.72
LSSLR	1.36	4.19	11.6
SLR	1.43	5.05	11.7

The results from the simulation studies are in Table 2. The optimal tuning parameters selected by the GIC in our models were $\gamma_1 = 1.00$, $\gamma_2 = 0.102$, and $\lambda = 10^{-2.50}$ for Case 1, $\gamma_1 = 1.00$, $\gamma_2 = 0.106$, and $\lambda = 10^{-1.98}$ for Case 2, and $\gamma_1 = 1.00$, $\gamma_2 = 0.106$, and $\lambda = 10^{-1.98}$ for Case 3, respectively. From the simulation results, we observe that our proposed procedure performs well in all cases with respect to minimizing prediction error rates even though Case 2 is an ordinary setting of semi-supervised learning, i.e., the density function for labeled data is same as that for unlabeled data. Hence, we conclude that our proposed method may be useful even if the densities for labeled and unlabeled data are same.

4.3 Benchmark data analysis

Thorough analyzing g10 data set (Chapelle and Zien, 2005), ionosphere data set (Sigillito *et al.*, 1989), and pima data set (Ripley, 1996), we illustrated the effectiveness of the proposed semi-supervised methodology. The g10 data set includes 550 data points with 10 predictors, and we prepared 250 training data points and 300 test data points. The ionosphere data set consists of 356 data points with 33 predictors, and we split the whole 356 data points into 150 training data points and 201 test data points. The pima data set, which consists of 300 training data points and 232 test data points, is a binary classification with 7 predictors. In order to implement semi-supervised procedure, the training data points were randomly split into two halves with labeled data points and unlabeled data points, where labeled data points were assigned as 5%, 10%, 20%, 30%, 40%, and 50% for training data points, respectively. We repeated the random splitting

Table 3: Comparisons of prediction error rates (%) for some data sets.

Method \ %	5	10	20	30	40	50
g10						
SSLRCS	3.40	3.47	3.85	4.06	4.66	5.42
LSSLR	26.6	16.2	9.94	7.04	5.66	4.77
SLR	26.4	16.4	9.30	6.85	5.45	4.62
Ionosphere						
SSLRCS	18.2	17.3	16.9	16.4	17.3	16.8
LSSLR	29.0	22.8	18.9	17.4	16.2	15.4
SLR	28.9	23.1	19.5	18.0	16.7	15.7
Pima						
SSLRCS	26.6	26.9	26.6	26.8	26.7	26.7
LSSLR	30.1	27.0	27.0	27.0	26.9	26.7
SLR	29.3	26.9	26.9	27.0	26.8	26.7

50 times. We also compared our proposed method (SSLRCS) with the LSSLR and the SLR, which is described in Section 4.1.

Table 3 shows the prediction errors for the benchmark data sets. We obtained optimal values of tuning parameters included in our proposed models as follows: $\gamma_1 = 1.00$, $\gamma_2 = 0.154$, and $\lambda = 10^{-3.20}$ for g10 data set, $\gamma_1 = 0.992$, $\gamma_2 = 0.504$, and $\lambda = 10^{-2.89}$ for ionosphere data set, and $\gamma_1 = 1.00$, $\gamma_2 = 0.308$, and $\lambda = 10^{1.41}$ for pima data set, which are averaged over 50 repetitions. From the results, we find that our proposed procedure outperforms the previously proposed methods in almost all situations, although it is unclear that whether densities for labeled and unlabeled data are different. In particular, the proposed method seems to work well when the number of labeled data points is small.

5 Concluding remarks

We proposed a semi-supervised logistic classification methodology for different density functions of labeled and unlabeled data along with the technique of covariate shift adaptation and regularization. A crucial point for our semi-supervised modeling processes includes the choices of some tuning parameters in our proposed models. We introduced a model selection criterion from the viewpoints of information-theoretic approach in order to select the values of the adjusted parameters. Through Monte Carlo simulations and benchmark data analysis, we showed that our modeling strategy is effectiveness in practical situations in the viewpoints of yielding relatively lower prediction errors than previously developed methods. Our modeling procedure may be applied into the problem of constructing a nonlinear semi-supervised discriminant model based on basis expansion, which will be discussed in another paper.

References

- [1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions Automatic Control*, **AC-19**, 716–723.
- [2] Amini, M-R. and Gallinari, P. (2002). Semi-supervised logistic regression. *Proceedings of the 15th European Conference on Artificial Intelligence*, 390–394.
- [3] Ando, T., Konishi, S. and Imoto, S. (2008). Nonlinear regression modeling via regularized radial basis function networks. *Journal of Statistical Planning and Inference*, **138**, 3616–3633.
- [4] Basu, S., Bilenko, M. and Mooney, R. J. (2004). A probabilistic framework for semi-supervised clustering. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press, 59–68.
- [5] Bennett, K. P. and Demiriz, A. (1998). Semi-supervised support vector machines. *Advances in Neural Information Processing Systems*, **11**, 368–374.

- [6] Bickel, S., Brückner, M. and Scheffer, T. (2009). Discriminative learning under covariate shift. *Journal of Machine Learning Research*, **10**, 2137–2155.
- [7] Chakraborty, S. (2011). Bayesian semi-supervised learning with support vector machine. *Statistical Methodology*, **8**, 68–82.
- [8] Chapelle, O., Schölkopf, B. and Zien, A. (2006). *Semi-Supervised Learning*. Cambridge, MA: MIT Press.
- [9] Chapelle, O. and Zien, A. (2005). Semi-supervised classification by low density separation. *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, 57–64.
- [10] Chen, K. and Wang, S. (2007). Regularized boost for semi-supervised learning. *Advances in Neural Information Processing Systems*, **20**, 281–288.
- [11] Dean, N., Murphy, T. B. and Downey, G. (2006). Using unlabelled data to update classification rules with applications in food authenticity studies. *Journal of the Royal Statistical Society Series C*, **55**, 1–14.
- [12] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. 2nd ed. New York: Springer.
- [13] Huang, J., Smola, A., Gretton, A., Borgwardt, K. M. and Schölkopf, B. (2007). Correcting sample selection bias by unlabeled data. *Advances in Neural Information Processing Systems*, **19**, 601–608.
- [14] Jiang, J. and Zhai C-X. (2007). Instance weighting for domain adaptation in NLP. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 264–271.
- [15] Kanamori, T., Hido, S. and Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, **10**, 1391–1445.
- [16] Kawano, S. Misumi, T. and Konishi, S. (2010). Semi-supervised logistic discrimination via graph-based regularization. Preprint, MI2010-6, Kyushu University.

- [17] Kawano, S. and Konishi, S. (2011). Semi-supervised logistic discrimination via regularized Gaussian basis expansions. *Communications in Statistics - Theory and Methods*, **40**, 2412–2423.
- [18] Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, **83**, 875–890.
- [19] Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. New York: Springer.
- [20] Lafferty, J. and Wasserman, L. (2007). Statistical analysis of semi-supervised regression. *Advances in Neural Information Processing Systems*, **21**, 801–808.
- [21] Liang, F., Mukherjee, S. and West, M. (2007). The use of unlabeled data in predictive modeling. *Statistical Science*, **22**, 189–205.
- [22] Ripley, B. D. (1996): *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.
- [23] Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, **90**, 227–244.
- [24] Sigillito, V. G., Wing, S. P., Hutton, L. V. and Baker, K. B. (1989). Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, **10**, 262–266.
- [25] Sokolovska, N., Cappé, O. and Yvon, F. (2008). The asymptotics of semi-supervised learning in discriminative probabilistic models. *Proceedings of the 25th International Conference on Machine Learning*.
- [26] Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Büna, P. and Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, **60**, 699–746.

- [27] Vittaut, J-N., Amini, M-R. and Gallinari, P. (2002). Learning classification with both labeled and unlabeled data. *Proceedings of the 13th European Conference on Machine Learning*, 468–479.
- [28] Wu, D., Lee, W. S. and Ye, N. (2009). Domain adaptive bootstrapping for names entity recognition. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 1523–1532.
- [29] Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. *Proceedings of the 21th International Conference on Machine Learning*, 114–121.
- [30] Zhou, D., Bousquet, O., Lal, T. N., Weston, J. and Schölkopf, B. (2004). Learning with local and global consistency. *Advances in Neural Information Processing Systems*, **16**, 321–328.
- [31] Zhu, X. (2008). Semi-supervised learning literature survey. Computer Sciences Technical Report 1530, University of Wisconsin-Madison.
- [32] Zou, H., Zhu, J., Rosset, S. and Hastie, T. (2007). Automatic bias correction methods in semi-supervised learning. *Contemporary Mathematics*, **443**, 165–175.